Introduction to Data Science
Autumn 2018
Team :
Sophie Gräfnitz
Evelin Aasna
Marianne Dengo

22.11.2018

First Steps of the Project
KAGGLE-CLINICAL SURVEY

Task 1 Business Understanding

Identifying our Business Goals
This project is an obligatory component of the course *Introduction to Data Science*, in Autumn Semester 2018 at *University of Tartu*. The goal of the project is to apply the learned techniques independently and successfully on some self-chosen data set. Hence, the overall business goal is at the course instructors' side. We can only assume, that the goal is to educate the students and to make them pass the course with a big knowledge gain.

Assessing your situation
Inventory of Resources
Our resources comprise brain power of the team member, the team's private technical equipment and the equipment available to students of *University of Tartu* in general. Our time resource for the whole project is 26 days (calculated from 22nd of November) or 21 days after the project pitch on 28th of November. Regarding software, we will rely on the Data Science packages from *Python* programming language.

Requirements, Assumptions and Constraints
We assume our data set is relatively clean and minable with the given technologies from the lecture. Unfortunately, we are constrained in our computing resources. Furthermore, some kind of project schedule is needed and continuously reporting according to CRISP-DM to track our workflow and to detect errors.

Risks and Contingencies
As there are no kernels published yet on *kaggle.com* for this particular data set, there might be the case that it is not relevant enough or not of enough interest. However, this circumstance is very unlikely - and even if our insights will not be significant, they are still insights which will be sufficient for passing the course. The team members' schedule might prevent the team from meeting the deadline, but good time management should handle this.

Terminlogy
Terminology: Create a list of business terms and data-mining terms that

are relevant to your project and write them down in a glossary with definitions (and perhaps examples), so that everyone involved in the project can have a common understanding of those terms.
We will fulfill this over time.

Costs and Benefits
This project will cost no money, but it will cost the team's time and energy. Still, our benefits will be knowledge gain and some interesting insights into the Indian social community which might even help development organizations or the Indian administration. Finally, completing any kind of project will strengthen our competencies and satisfy our eagerness to create.

Defining our Data Mining Goals
Data Mining Goals
The main goal is to derive a reasonable hypothesis from the data set applying reasonable data mining methods. Meanwhile, the team needs to deliver the project code of the data preparation, modeling, evaluation and deployment and a summary of a project report as a presentation poster.

Data Mining Success Criteria
The success of our project depends on the instructors *Meelis Kull, Mikk Puustusmaa* and the rest of the *Introduction to Data Science* Team. The formal evaluation criteria were given in a project kickoff lecture. Every team member will receive the team grading of maximum 20 points. 10 points can be achieved by representational quality, which means our poster will hold all necessary content such as main results, applied data science methods, motivation and objectives. 10 more points can be reached by technical quality, which means to state a clear objective in our report, gain relevant insight from the chosen relevant data and execute everything time-efficiently.

Task 2 Data Understanding

Gathering Data
Outline data requirements
The data needs to be provided in a sklearn compatible format, such as csv.

Verify data availability
The data set *Clinical, Anthropometric & Bio-Chemical Survey* is publicly available on *kaggle.com* in csv-format. When importing the csv data sets, param *low_memory* = *True* is necessary as there are different type columns in the data sets.

Introduction to Data Science
Autumn 2018
Team :
Sophie Gräfnitz
Evelin Aasna
Marianne Dengo

22.11.2018

Define selection criteria

The only data source is *https://www.kaggle.com/rajanand/cab-survey* last called the 22.11.18 at 12.02h in Tartu, Estonia. We choose the biggest data set out of `'CAB_22_CT.csv', 'CAB_05_UT.csv', 'CAB_20_JH.csv', 'CAB_23_MP.csv', 'CAB_08_RJ.csv', 'CAB_data_dictionary.xlsx', 'CAB_21_OR.csv', 'CAB_10_BH.csv', 'CAB_09_UP.csv', 'CAB_18_AS.csv',so it is 'CAB_09_UP.csv'.`

Our goals is to find good features for each of the three disease indicators blood sugar, blood haemoglobin, and blood pressure. For each of these disease indicators, the feature selection (and sample selection) must be done differently:

- outcome of blood sugar testing (features 36-38)
  - fasting blood glucose level out of 70-100 mg/dL range indicates diabetes
  - only adults
- blood haemoglobin for individuals 6 months or older (features 27-29)
  - blood hemoglobin below 13.5g/dL in men or 12.0g/dL in women indicates anemia
- systolic/diastolic blood pressure (features 30-33)
  - only adults

After trying to predict/recognize patterns with these features, we can add other features that increase outcome.

Describing Data

We have Uttar Pradesh data set provided in csv format. The data set consists of approximately 490.000 cases and 53 features. The features can be assigned to either survey related data (such as date, state), general personal ID data (such as sex, age…) and individual health data (such as haemoglobin, pulse rate, illness type). Of the 53 features some are only relevant depending on the age/sex of the individual:

- features 22-26 for individuals aged 1 month or older
- features 27-29 for individuals aged 6 months or older
- features 30-38 for individuals aged 18 years or older
- features 39-41 for women aged 15-49
- features 42-50 for children under 3 years
- features 51-53 for children under 5 years

Depending on our goal we can probably discard several cases. There is data about 1.89 million individuals so the number of cases will probably still be sufficient. As the regarded Indian provinces have many inhabitants, in the end the sample size needs to be considered when drawing conclusions.

Introduction to Data Science
Autumn 2018
Team :
Sophie Gräfnitz
Evelin Aasna
Marianne Dengo

22.11.2018

Exploring Data

Initial exploration revealed that all states except for Rajasthan have approximately balanced number of male and female individuals. Based on the data dictionary attached to the dataset feature age_code marks if the age feature shows age in years(Y), months(M) or days(D). For Rajasthan data this column is numeric and not easily interpreted. These finding suggest we should exclude Rajasthan data.

Verifying Data Quality

Further exploration revealed some input errors that have to be dealt with. This should not become a major problem. Questioned individuals also had the option of refusing blood sugar/blood pressure/haemoglobin testing. We have looked at the proportion of individuals for whom testing was conducted. For each state and each indicator over 50% of total individuals consented to testing. This confirms that the data we need exists.

Introduction to Data Science
Autumn 2018
Team :
Sophie Gräfnitz
Evelin Aasna
Marianne Dengo

22.11.2018

Task 3 Setting Up and Planning the Project

Repository and Registration

Link to our project repository:
https://github.com/mdengo/cab-survey

Link to our project slide in the list of all projects:
https://docs.google.com/presentation/d/1RHDUPsJVVtwVfPp8-WxsK8udEpYOmf4Ki9NjtbgpLDU/edit#slide=id.g48274606ac_127_5